KIND: an Italian Multi-Domain Dataset for Named-Entity Recognition

Teresa Paccosi[1,2], Alessio Palmero Aprosio[1]

[1] Fondazione Bruno Kessler - [2] Università di Trento - {tpaccosi, aprosio}@fbk.eu

Background and motivation

Named-entity recognition (NER) is the NLP task which consists in identifying and classifying mentions of entities in texts, belonging to a set of predefined categories among which people, locations, and organizations are the most common.

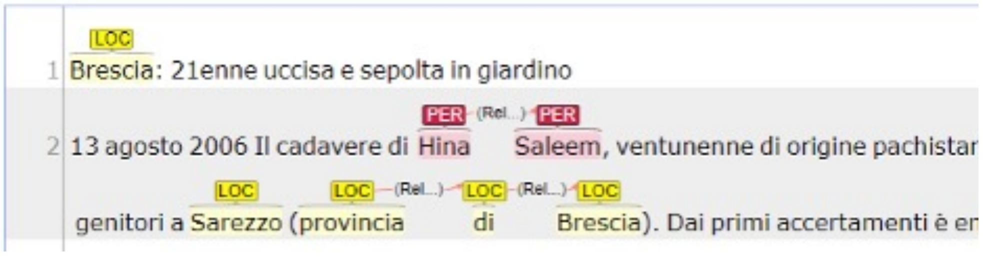
In our paper, we present KIND, Kessler Italian Named-entities Dataset, which is the largest Italian NER dataset with manual gold annotations, and it is the first Italian NER dataset which is multidomain, containing annotations of news articles, literary texts and political discourses.

KIND composition

- **Wikinews** - 1000 news articles (308.622 tokens)
- **Fiction** - 86 chapters from novels, epistles and biographies (192.448 tokens)
- **Aldo Moro** - 250 writings by Aldo Moro (392.604 tokens) (Barzaghi and Paolucci, 2021)
- **Alcide de Gasperi** - 158 public documents (150.632 tokens) (Tonelli et al., 2019)

Although the entities in Aldo Moro's set are already annotated, we performed a semi-automatic check, since their annotation guidelines are slightly different (*silver data*).

Annotation process and tagging scheme



1 Brescia: 21enne uccisa e sepolta in giardino

2 13 agosto 2006 Il cadavere di Hina Saleem, ventunenne di origine pachistar

genitori a Sarezzo (provincia di Brescia). Dai primi accertamenti è er

- Data preprocessing with TINT (Palmero Aprosio et al., 2018)
- Annotation tool: INCEpTION (Klie et al., 2018)
- IOB tagging scheme
  [Sophia]$_{B\text{-}PER}$ [Loren]$_{I\text{-}PER}$ is a famous italian actress

PER

Individuals and groups of people (including family names), or animals if they have a proper name.

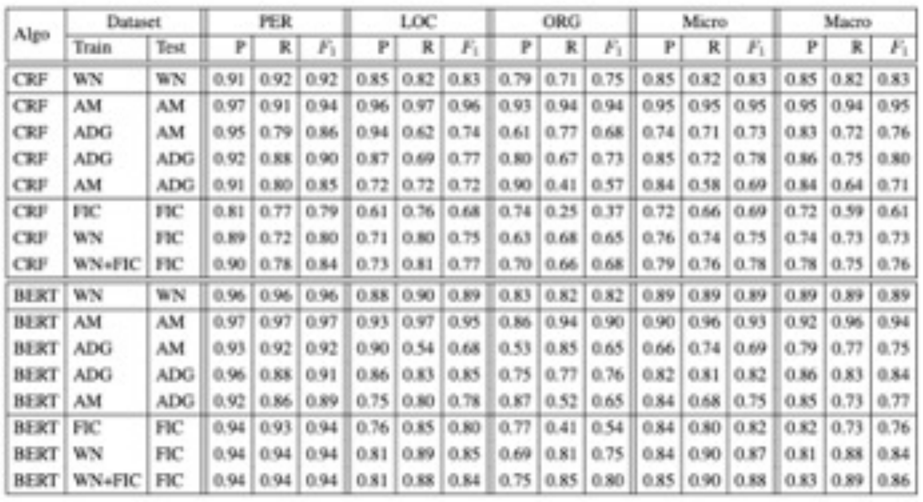
ORG

ORG entity must be a formally established association (governments, commercial organizations, media, etc.).

LOC

Places defined on a geographical or astronomical basis which are mentioned in a document and do not constitute a political entity.

Experiments and evaluation



Algo	Dataset		PER			LOC			ORG			Micro			Macro		
	Train	Test	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
CRF	WN	WN	0.91	0.92	0.92	0.85	0.82	0.83	0.79	0.71	0.75	0.85	0.82	0.83	0.85	0.82	0.83
CRF	AM	AM	0.97	0.91	0.94	0.96	0.97	0.96	0.93	0.94	0.94	0.95	0.95	0.95	0.95	0.94	0.95
CRF	ADG	AM	0.95	0.79	0.86	0.94	0.62	0.74	0.61	0.77	0.68	0.74	0.71	0.73	0.83	0.72	0.76
CRF	ADG	ADG	0.92	0.88	0.90	0.87	0.69	0.77	0.80	0.67	0.73	0.85	0.72	0.78	0.86	0.75	0.80
CRF	AM	ADG	0.91	0.80	0.85	0.72	0.72	0.72	0.90	0.41	0.57	0.84	0.58	0.69	0.84	0.64	0.71
CRF	FIC	FIC	0.81	0.77	0.79	0.61	0.76	0.68	0.74	0.25	0.37	0.72	0.66	0.69	0.72	0.59	0.61
CRF	WN	FIC	0.89	0.72	0.80	0.71	0.80	0.75	0.63	0.68	0.65	0.76	0.74	0.75	0.74	0.73	0.73
CRF	WN+FIC	FIC	0.90	0.78	0.84	0.73	0.81	0.77	0.70	0.66	0.68	0.79	0.76	0.78	0.78	0.75	0.76
BERT	WN	WN	0.96	0.96	0.96	0.88	0.90	0.89	0.83	0.82	0.82	0.89	0.89	0.89	0.89	0.89	0.89
BERT	AM	AM	0.97	0.97	0.97	0.93	0.97	0.95	0.86	0.94	0.90	0.90	0.96	0.93	0.92	0.96	0.94
BERT	ADG	AM	0.93	0.92	0.92	0.90	0.54	0.68	0.53	0.85	0.65	0.66	0.74	0.69	0.79	0.77	0.75
BERT	ADG	ADG	0.96	0.88	0.91	0.86	0.83	0.85	0.75	0.77	0.76	0.82	0.81	0.82	0.86	0.83	0.84
BERT	AM	ADG	0.92	0.86	0.89	0.75	0.80	0.78	0.87	0.52	0.65	0.84	0.68	0.75	0.85	0.73	0.77
BERT	FIC	FIC	0.94	0.93	0.94	0.76	0.85	0.80	0.77	0.41	0.54	0.84	0.80	0.82	0.82	0.73	0.76
BERT	WN	FIC	0.94	0.94	0.94	0.81	0.89	0.85	0.69	0.81	0.75	0.84	0.90	0.87	0.81	0.88	0.84
BERT	WN+FIC	FIC	0.94	0.94	0.94	0.81	0.88	0.84	0.75	0.85	0.80	0.85	0.90	0.88	0.83	0.89	0.86

Before running the experiments to train the models for the task of NER, we split the different datasets in train and test.

Fiction dataset. We used the works of 2 authors (Fabiani and Pavese) as test (not included in the train set) to avoid training the model on a writing style.

CRF model. Word shapes, 6-grams, previous, current, and next token/lemma/class.

BERT model. Token classification head on top. The model is trained (3 epochs) starting from the bert-base-italian-cased model.

Main issue in silver data (Aldo Moro's works)

All the locations in AM are tagged as LOC even if they refer to an organization. The most common cases are:

- Sport teams ([Italy] have won the gold medal)
- Governmental bodies ([Germany] withdrew from negotiations)

Final remarks and future directions

- Main strengths of KIND is the multidomain feature and the great amount of manual gold annotations.
- Dataset download: https://github.com/dhfbk/KIND
- We plan to add more documents especially in the fiction set and all the texts from Aldo Moro's work.